Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.35944

Text-to-Image Generation Based on Multimodal Fusion

Tianrui Shi

Department of Network Engineering, Beijing Information Science and Technology University, Beijing, China
2023011483@bistu.edu.cn

Abstract. Text-to-image (T2I) generation refers to the technology that takes natural language descriptions as input and automatically generates images semantically aligned with them, which is one of the core tasks in the field of cross-modal content generation. However, existing technologies still face core challenges in handling complex semantics, restoring fine details, and balancing generation diversity and controllability, making it difficult to fully meet the complex needs of diverse application scenarios relying solely on text prompts. To address these challenges, generation methods can be designed by improving multimodal fusion strategies based on different fusion approaches. For this purpose, from the perspective of multimodal fusion and with fusion timing as the main thread, this paper systematically sorts out mainstream T2I technologies into four main paradigms: early fusion, late fusion, attention mechanism-based fusion, and hybrid fusion. It systematically explains their core mechanisms, evolution trends, and conducts model performance comparisons. The paper further sorts out the key datasets supporting the development of the field, aiming to provide help and inspiration for subsequent related research in the field, identify potential research directions, and promote the practical application and innovative breakthroughs of text-to-image generation technology in more complex scenarios.

Keywords: Multimodal Fusion, T2I, AIGC

1. Introduction

Research on text-to-image generation has undergone a significant paradigm shift from Generative Adversarial Networks to diffusion models, with its core consistently lying in exploring more efficient and accurate multimodal fusion methods. Early research was dominated by Generative Adversarial Networks, whose core idea is to learn data distribution through adversarial training between generators and discriminators. By contrast, diffusion models, via a stable gradual denoising process, significantly outperform Generative Adversarial Networks (GANs) in terms of generation quality and diversity. This technological evolution has laid a solid foundation for meeting current generation requirements for open-domain and complex-scenario tasks, and also makes designing fusion strategies to fully unleash the potential of models one of the most core research topics in this field.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.35944

Nevertheless, current models still face three core challenges: first, biased understanding of complex descriptions; second, insufficient fine-grained restoration; third, the difficulty in balancing controllability and generation diversity. The core to solving these problems lies in multimodal fusion strategy, since the design of the fusion strategy directly affects the model's ability to understand and follow text prompts [1]. The timing of fusion acts as a distinct control variable in the continuous generation process, which directly defines the depth, frequency and scope of information exchange between modalities, and further affects the overall performance of the model. A classification framework based on fusion timing can intuitively reflect the impact of different fusion strategy designs on model performance. To this end, this paper categorizes existing fusion strategies according to fusion timing, namely early fusion, late fusion, attention-based fusion and hybrid fusion. Following this main line, this paper systematically reviews and compares representative models of various strategies, analyzes their mechanisms, advantages and limitations, and reveals the evolution trend of this field. The subsequent chapters of this paper are organized in accordance with this context.

2. Early fusion

2.1. Concept and characteristics of early fusion

Early fusion, also known as front-end fusion, is one of the most intuitive strategies in multimodal fusion. Its core idea is generated in the process of the front, that is, the input layer or shallow network, from different modal characteristics of one-time, static integration, forming a unified representation, and then passed on to the subsequent network for processing and transformation. In the text-to-image generation task, before the Generator starts to work, the semantic information of the text is bound with the latent variables that determine the content and structure of the image (such as random noise), and then input into a unified model for processing [2].

Using the model of this strategy in the generation process of the front, the text information is fused with random noise, and then the input is fed into the generation network for forward calculation results; the follow-up process processes the text and image characteristics without dynamic interaction. Therefore, this design makes the generation process have a clear global goal from the beginning, which is conducive to quickly determining the overall composition, main body and tone of the image. It is highly efficient in generating images with relatively simple structure and clear theme; its training is relatively stable, and the computational overhead is low. However, due to the lack of feedback and the unidirectional transmission of text information, the model is easy to lose or confuse the details in long texts and complex descriptions. If the coding is not accurate enough in the "text encoder" link, or information loss is generated during fusion, this deviation will run through the entire generation process without correction opportunity and be amplified by subsequent network layers.

2.2. Text image generation based on early fusion

Common early fusion models include GAN-INT-CLS, StackGAN, StackGAN++, etc. The common characteristics of these models are a simple and intuitive structure, fewer training process steps, but complex semantic description follows the poor and inadequate reduction of fine-grained [3].

Of GAN - INT - CLS will be generated for the first time against network combined with the text characteristic encoder, to verify using against the feasibility of training directly from the text to

DOI: 10.54254/2755-2721/2026.35944

generate images, model using the simple way to early fusion, is foundation of early fusion, but its only can generate low resolution images, fuzzy and details.

StackGAN's core idea is to use two phase structure will be difficult to high resolution image generated task is decomposed into an outline sketch with details elaboration, two substages are easier to handle. StackGAN also proposes the conditional augmentation technique. It maps text features to the parameters of a Gaussian distribution and samples conditional variables from that distribution. This increases the smoothness and diversity of the condition space, which is then effectively stabilized by the KL divergence regularization.

StackGAN++ is an improved model based on StackGAN. It uses an end-to-end tree-structured multiple generator-multiple discriminator architecture to jointly approximate the image distribution at multiple scales, thus achieving more stable training and higher quality generation. The backbone parameters are shared by all generators, and the joint training design enables the model to simultaneously learn the multi-scale image distribution from low to high. Low-scale generation provides a good foundation for high-scale generation, and the gradient generated at high-scale can reverse optimize the low-scale features, forming a positive feedback loop, which significantly improves the training stability and the final output quality.

The performance comparison data information of each early fusion model IS shown in Table 1. The test set is CUB-200-2011, and the indicators include IS and FID.

Table 1. Performance comparison of early fusion models

Fusion strategy	Model name	Year of publication	Test set	Test resolution	IS↑	FID↓
Early fusion	GAN-INT-CLS	2016	CUB-200-2011	64 x 64	2.88 ± 0.04	68.79
Early fusion	StackGAN	2017	CUB-200-2011	64 x 64	3.7	51.89
Early fusion	StackGAN-v2	2018	CUB-200-2011	128 x 128	4.04 ± 0.05	15.30

3. Late fusion

3.1. Concept and characteristics of late fusion

Late fusion, also known as back-end fusion, is another important technical path in multimodal fusion. The core idea of late fusion is the opposite of early fusion. It introduces text information at the end of the process or through an external loop to guide, filter, or optimize the generated or nearly formed image content.

In the text-to-image generation task, the late fusion model focuses on generating visually reasonable or diverse images at the beginning, and the text description is used as a post-hoc constraint or optimization goal to evaluate the quality and matching degree of the generated results and adjust them. Behind the design train of thought is to first ensure the visual fidelity and diversity of the image itself, and then with the help of the text, it constraint in semantic on the right track. This structure makes the model using this strategy not be constrained by strong text conditions in the early generation stage, which is conducive to exploring a broader visual possibility space, and then generating more diverse images.

Late fusion is often combined with pre-trained cross-modal models such as CLIP to exploit strong external priors. These models are trained on huge Internet data and have powerful image-text matching judgment ability, which can be used as efficient and training-free external guiders to give optimization directions for any generative model.

DOI: 10.54254/2755-2721/2026.35944

3.2. Based on the late fusion image text generation

Common models of late fusion include MirrorGAN, LAFITE, StyleGAN-T, etc. The common feature of these models is that they use decoupled generation, and all rely on external or internal alignment mechanisms to achieve back-end alignment. The generated images have high quality, good visual diversity and semantic consistency, and have flexibility and potential in solving problems such as semantic alignment, data efficiency and architecture scalability. However, due to the late introduction of text, the generation process is less controllable, and the ability to deal with spatial relations, multi-attribute binding, and local modification is weak. The late fusion model can be further classified into two types according to how the text is used to evaluate and optimize the generated image.

The first is architecture-based end-to-end training, which is a strategy to internalize the idea of late alignment into the model architecture, aiming to make the model learn to "self-reflect" through training. MirrorGAN establishes a "text → image → text" mirroring closed loop, which performs self-supervision [4] through image re-description. After generating an image, it is re-described as text by a pre-trained image caption generator. The semantic reconstruction loss between the original text and the re-described text is calculated, and the gradient is back-propagated to the generator to force the generator to produce semantically consistent images.

The second is external guidance based on optimization, which is a model-independent late fusion strategy that does not modify the internal structure of the generator. Taking CLIP-Guided as an example, after the base generative model generates an initial image I-0 from random noise, the pre-trained CLIP model is used to encode the image I-0 and the target text description t, respectively, and the negative value of the cosine similarity between them in the CLIP joint feature space is calculated. Finally, through gradient descent, the gradient of the loss with respect to image pixel I is calculated and the image is iteratively updated by this gradient. This process is repeated so that the generated image is gradually closer to the text description in CLIP space. LAFITE only leverages the image set and the powerful cross-modal alignment prior of the pre-trained CLIP model to learn text-to-image generation [5] to achieve training without paired data acquisition. The model completely removes the text encoder, and only the features of the input text obtained by the CLIP text encoder are directly input into the generator as the target conditions during inference. The model uses the CLIP prior to get rid of the constraints of text data and achieve efficient feature space alignment. Based on StyleGAN - XL StyleGAN -t is able to adapt to more modal data, with large capacity and stable and can weigh the generated by precision control to balance diversity and text alignment, a dedicated GAN architecture. Through systematic architecture remodeling and external model guidance, this model pushes the performance of GAN to new heights on large datasets.

The performance comparison data information of each late fusion model IS shown in Table 2. The test set is MS-COCO (30K), and the indicators include IS, FID, R-score, and CLIP Score.

Table 2. Performance comparison of late fusion models

Fusion strategy	Model name	Year of publication	Test set	Test resolution	IS↑	FID↓	R-precision↑	CLIP Score↑
Late fusion	MirrorGAN	2019	MS-COCO (30K)	128 x 128	26.47 ±.41	-	74.52	-
Late fusion	Lafite	2021	MS-COCO (30K)	256 x 256	26.02	8.12	-	0.336

DOI: 10.54254/2755-2721/2026.35944

Table 2. (continued)

Late fusion	StyleGAN-T	2023	MS-COCO (30K)	256 x 256	-	2.10	-	0.340

4. Fusion based on attention mechanism

4.1. Fusion concepts and characteristics based on attention mechanism

The fusion based on the attention mechanism is also called mid-term fusion. The fusion opportunity is in the middle layer of the whole process of generation, and the dynamic depth interaction between text and image is realized by introducing the attention mechanism. The attention mechanism gives the model the ability to dynamically focus and associate information, which fundamentally solves the key problems in sequence modeling and cross-modal alignment. It has become a mainstream solution to achieve high-quality and high semantic controllability generation, and is one [6] of the common modules of modern models.

4.2. Text image generation based on attention mechanism

Common models of fusion based on the attention mechanism include AttnGAN, Stable Diffusion1.5, DALL-E, SDXL, etc. The common characteristics of these models are that they adopt the cross-attention mechanism as the core technology, have the ability to fine-regulate text in the image generation process, and can effectively solve complex semantic generation tasks. The continuous development and optimization of the core mechanism make T2I technology break through the bottleneck, constitute the backbone of the current technology and promote the development of the whole field.

Among them, AttnGAN demonstrates for the first time that the attention mechanism can enable GAN to achieve high-level fine-grained generative control. The concept of attention generation network and DAMSM loss is proposed, which lays the foundation for subsequent attention-based fusion models and is the key turning point to promote the leap of T2I generation quality.

Stable Diffusion1.5 is an open-source T2I model [7] released by Stability AI based on LDM. This model has improved image quality, generation stability, and the ability to follow complex prompt words. It has quickly become the most popular and ecologically prosperous basic model in the open source community, and is a landmark practice of attention-based fusion strategy in the open source field. It achieves efficient and high-quality text-conditional image generation by deeply integrating cross-attention into the UNet of the latent diffusion model, and promotes the rapid development of the entire AIGC field with its openness. Its performance and limitations also point the way for the improvement of next-generation models (e.g., SDXL, SD3). SDXL takes cross-attention-based fusion to the next [8]level by comprehensively upgrading the model size, conditional encoding, and training strategy.

Although DALL-E does not focus on improving the attention mechanism itself, it verifies whether a unified, transformer-based autoregressive architecture can directly learn the joint distribution of text and image after absorbing massive image-text pairs, and achieve high-quality zero-shot generation.

The performance comparison data information of each fusion model based on the attention mechanism IS shown in Table 3. The test set is MS-COCO (30K), and the indicators include IS, FID, and CLIP Score.

DOI: 10.54254/2755-2721/2026.35944

Table 3. Performance comparison of fusion models based on attention mechanism

Fusion strategy	Model name	Year of publication	Test set	Test resolution	IS↑	FID↓	CLIP Score↑
Attention-based fusion	AttnGAN	2018	MS-COCO (30K)	256 x 256	25.89	~ 24.76	-
Attention-based fusion	Stable Diffusion 1.5	2022	MS-COCO (30K)	256 x 256	-	12.5	0.312
Attention-based fusion	DALL-E 2	2022	MS-COCO (30K)	256 x 256	-	10.39	0.32
Attention-based fusion	SDXL	2023	MS-COCO (30K)	1024 x 1024	-	8.9	0.31

5. Mix and blend

5.1. Concept and characteristics of hybrid fusion

Hybrid fusion has evolved into a system-level architecture design idea. The core goal of hybrid fusion is no longer a simple timeline splicing, but according to task requirements, intelligent scheduling and fusion of multiple modal interaction mechanisms, including early and late thoughts, attention mechanisms, external model priors, specific functional modules, etc., to achieve a comprehensive ability that cannot be achieved by a single strategy. Therefore, hybrid fusion systematically enhances the controllability and ultimate quality while retaining the powerful generation ability of the base model by introducing parallel control branches and concatenating dedicated optimization stages.

5.2. Introduction of hybrid fusion model

Common models of hybrid fusion include XMC-GAN, UniDiffuser, SD3, Flux.1, etc. The common feature of these models is that they cleverly combine multiple fusion strategies, network modules and training objectives, and are no longer limited to a single time to pursue extreme performance and controllability. Hybrid thinking is the current mainstream and evolution trend. Although it has excellent comprehensive performance, it also has the highest computational cost.

By introducing cross-modal contrastive learning as an additional supervision signal, the XMC-GAN model is combined with the attention-based generation mechanism to generate images that are not only visually realistic but also highly semantically consistent with the text description. On the basis of the classical attention generation architecture, the model also innovatively adds a cross-modal contrastive loss, which realizes the strategy mix of discriminative alignment and generative modeling.

UniDiffuser is a diffusion architecture that unifies multiple generative tasks. This architecture explores the general potential of diffusion models on multi-modal tasks, enabling it to handle various joint and conditional generation tasks of images and text within a single framework, enabling a mixture of modalities and task flows.

The SD3 model innovatively designs a framework to replace the traditional U-Net with a multi-modal diffusion Transformer, which achieves more accurate text understanding and image generation [9] through the powerful sequence modeling ability of the pure Transformer architecture. The SD3 model combines two powerful text encoders: CLIP and T5-XXL. CLIP provides excellent

DOI: 10.54254/2755-2721/2026.35944

visual semantic alignment priors, while T5-XXL provides strong language understanding with long text encoding capabilities. At the same time, the outputs of the two are fused at the early stage of the model to provide richer information conditions for the subsequent diffusion Transformer.

Flux.1, introduced in recent studies, represents the pinnacle of current hybrid fusion strategies. Its goal is to achieve the ultimate generation quality, prompt compliance, and reasoning ability [10] through an unprecedented systematic and large-scale integration of multiple dimensions such as data, architecture, and training objectives. This model achieves the SOTA of comprehensive performance without significantly sacrificing any dimensions.

The performance comparison data information of each hybrid fusion model IS shown in Table 4. The test set is MS-COCO (30K), and the indicators include IS, FID, and CLIP Score.

Table 4. Performance comparison of hybrid fusion models

Fusion strategy	Model name	Year of publication	Test set	Test resolution	IS↑	FID↓	CLIP Score↑
Mix fusion	XMC-GAN	2021	MS-COCO (30K)	128 x 128	9.33	9.3	-
Blended fusion	Diffuser	2023	MS-COCO (30K)	256 x 256	-	13.7	0.293
Blended fusion	SD 3	2024	MS-COCO (30K)	1024 x 1024	-	2.48	0.344
Blended fusion	Flux.1	2025	MS-COCO (30K)	1024 x 1024	-	1.85	0.348

6. Dataset

In the task of text-to-image generation based on multi-modal fusion, datasets are not only benchmarks for technical evaluation, but also important tools to drive the development of the field, focus on challenges, and measure the effectiveness of fusion strategies. Each evolution of T2I technology and innovation in model technology is inseparable from the data foundation on which it is trained and evaluated.

At present, the main datasets used in text-generated image research are CUB-200-2011, Oxford-102 Flowers, Multi-modal-CelebA-HQ, MS COCO, etc [11].

The landscape of text-to-image (T2I) generation research is underpinned by a collection of specialized and general-purpose datasets, each serving distinct evaluative and developmental functions.

Foundational work in fine-grained generation is often benchmarked using the CUB-200-2011 and Oxford-102 Flowers datasets, introduced in 2011 and 2008, respectively. These relatively compact collections, containing 11,788 avian images across 200 species and 8,189 images of 102 flower categories, are characterized by their detailed text annotations focused on subtle, discriminative attributes such as beak morphology or petal texture, thereby enabling a rigorous assessment of a model's capacity to render specific local features with high fidelity.

For high-resolution human-centric synthesis, the MM CelebA-HQ dataset, established in 2020, supplies 30,000 facial images at a resolution of 1024×1024, augmented with natural language descriptions designed specifically to test multimodal control and precise attribute manipulation.

General scene comprehension and zero-shot generalization capabilities are predominantly trained and evaluated on the ubiquitous MS COCO dataset, a corpus of over 120,000 complex scenes initiated in 2014, where each image is paired with five manual captions describing multiple objects and their interrelations.

The paradigm shift towards large-scale pre-training was fundamentally enabled by web-scale datasets, notably the LAION-5B corpus, a vast collection of approximately 5.85 billion image-text pairs released in 2022 that, despite its variable quality, serves as the core data source for

DOI: 10.54254/2755-2721/2026.35944

foundational models like Stable Diffusion. An earlier precedent for this approach was the Conceptual Captions (CC12M) dataset, which provided around 12 million concise, colloquial alt-text pairs and facilitated early research into learning semantics from weakly supervised data.

Beyond image-text pairings, the field's progress in compositional generation is measured using prompt-only benchmarks. The 2022 DrawBench suite, with roughly 200 systematically crafted prompts, exposes model deficiencies in combinatorial reasoning involving color, spatial relationships, and object counting. This was subsequently scaled up by the T2I-CompBench benchmark in 2023, which offers a standardized toolkit of approximately 6,000 prompts to deliver quantifiable scores across subtasks like attribute binding, enabling systematic, fine-grained performance comparisons between models.

7. Conclusion

Taking fusion timing as the core thread, this paper systematically reviews four mainstream multimodal fusion strategies in the field of text-to-image generation, namely early fusion, late fusion, attention-based fusion, and hybrid fusion. Their evolution is a rising process from simple injection to system integration by constantly breaking through the core bottleneck. This paper provides a clear and coherent cognitive map for understanding the transition in this field and provides a clear theoretical perspective for understanding the evolution logic of multimodal generative models by establishing an interpretation mode centered on fusion timing. This study is expected to provide useful reference and inspiration for the subsequent exploration of more efficient, intelligent, and controllable cross-modal content generation technology.

References

[1] Li, S., & Tang, H. (2025). Multimodal alignment and fusion: A survey. *International Journal of Computer Vision*. https://doi.org/10.48550/arXiv.2411.17040

[2] Li, C., & Wu, H. (2025). Research on image generation methods based on generative adversarial networks. *Information & Computer, 37*(10), 1–5.

[3] Zhang, S., Chen, X., & Liu, Q. (2022). Text-to-image generation with generative adversarial networks: A survey. *ACM Computing Surveys, 55*(6), 1–36. https://doi.org/10.1145/3544548

[4] Qiao, T., Zhang, J., Xu, D., et al. (2019). MirrorGAN: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1505–1514). IEEE.

[5] Zhou, Y., Zhang, R., Chen, C., et al. (2022). LAFITE: Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 15607–15617). IEEE.

[6] Yu, K. (2025). *Research on text-guided image generation methods based on adversarial networks* [Dissertation, Xi'an Shiyou University].

[7] Lin, Z., Chen, M., Luo, Y., et al. (2026). A review of controllable text-to-image generation techniques driven by visual priors. *Journal of Sichuan University of Science & Engineering (Natural Science Edition)*. Advance online publication. https://link.cnki.net/urlid/51.1792.n.20260522.0922.002

[8] Cao, P., Zhou, F., Song, Q., et al. (2025). Controllable generation with text-to-image diffusion models: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2025.3646548

[9] Stability AI. (2024, March 5). Stable Diffusion 3: Technical report [Technical report]. *arXiv*. https://arxiv.org/abs/2403.03206

[10] Black Forest Labs. (2025). FLUX.1: A foundational large model for text-to-image generation [Technical report].

[11] Wang, Q., Jiang, G., Weng, N., et al. (2026). A review of efficient image generation techniques based on generative adversarial networks. *Computer Science*. Advance online publication. https://link.cnki.net/urlid/50.1075.tp.20260526.1655.002 2